NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Lost in Translation, European Parliamentary Debates Transcripts

Alberto Parenti

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Lost in Translation, European Parliamentary Debates Transcripts

by

Alberto Parenti

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Prof Doutor Flávio Luís Portas Pinheiro

December, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Alberto Parenti

Lisbon, December 4, 2023

Acknowledgements

Thank you, Flavio, for your invaluable assistance with a project I am truly passionate about and for your endless patience. Your dedication to your work is inspiring.

Grazie Mamma for enabling me to study abroad, exposing me to diverse cultures, and for various other opportunities you have made possible over the years, all while making numerous sacrifices for a future we both dreamt of, which is now gradually taking shape thanks to you.

To all my friends, particularly Ivan and José, who were with me during my Master's journey in Portugal, and Federico and Lorenzo, whom I left for this adventure, I felt your moral support despite our rare meetups or being in different countries.

A huge thank you Filipa for your belief in my abilities and for your willingness to prioritize my needs above all else, offering your time and sharing ideas, making this achievement possible in ways I could not have managed alone.

”

“A different language is a different vision of life.”

— Federico Fellini

Abstract

This thesis explores the expression of emotions in formal speeches within the European Parliament (EP), analyzing the European Parliament Proceedings Parallel Corpus (EPPPC) that includes speeches from 1996 to 2011 in 21 languages.

Linguistic tools such as spaCy's monolingual and multilingual models are used for analyzing polarity and subjectivity, along with the multilingual Bidirectional Encoder Representations from Transformers (BERT) model for polarity assessment. It aims to identify potential changes in translation that might affect the emotional content of speeches.

This study uncovers notable sentiment biases with spaCy's multilingual model and observes substantial polarity and subjectivity shifts in language translation across various languages. These findings highlight how translation can profoundly alter the tone and sentiment perceived in political speeches.

Keywords: European Parliament, Language Translation, Sentiment Analysis, Natural Language Processing

Sustainable Development Goals (SDG):

Resumo

Esta tese explora a expressão de emoções em discursos formais no Parlamento Europeu, analisando o European Parliament Proceedings Parallel Corpus (EPPPC), que inclui discursos de 1996 a 2011 em 21 idiomas.

Ferramentas linguísticas como os modelos monolíngues e multilíngues do spaCy são usadas para analisar a polaridade e a subjetividade, juntamente com o modelo Bidirectional Encoder Representations from Transformers (BERT) multilíngue para avaliação da polaridade. O objetivo é identificar potenciais mudanças na tradução que possam afetar o conteúdo emocional dos discursos.

Este estudo revela enviesamentos significativos de sentimento com o modelo multilíngue do spaCy e observa mudanças substanciais de polaridade e subjetividade na tradução entre vários idiomas. Estes resultados destacam como a tradução pode alterar profundamente o tom e o sentimento percepcionados em discursos políticos.

Palavras-chave: Parlamento Europeu, Tradução de Línguas, Análise de Sentimentos, Processamento de Linguagem Natural

Sustainable Development Goals (SDG):

Contents

List of Figures

List of Tables

Acronyms

BERT Bidirectional Encoder Representations from Transformers *(pp. iii, iv, 4, 11, 12, 15–17, 22–27)*
BG Bulgarian *(pp. 6, 9)*
Bi-LSTM Bi-directional Long Short Term Memory *(p. 11)*

CNN Convolutional Neural Networks *(p. 11)*
CS Czech *(pp. 6, 9)*

DA Danish *(pp. 6, 9)*
DE German *(pp. 6, 9)*

EF European Foundation *(p. 8)*
EL Greek *(pp. 6, 9)*
EN English *(pp. 6, 7, 9)*
EP European Parliament *(pp. iii, 1, 2, 6, 7, 12, 13, 15, 25–27)*
EPPPC European Parliament Proceedings Parallel Corpus *(pp. iii, iv, 6–8, 14)*
ES Spanish *(pp. 6, 9)*
ET Estonian *(pp. 6, 9)*
EU European Union *(pp. 1, 3, 4, 6, 7, 26)*

FI Finnish *(pp. 6, 9)*
FR French *(pp. 6, 9)*

HU Hungarian *(pp. 6, 9)*

IT Italian *(pp. 6, 9)*

LSTM Long Short Term Memory *(p. 11)*
LT Lithuanian *(pp. 6, 9)*

LV Latvian *(pp. 6, 9)*

MEP Member of the European Parliament *(p. 1)*
MT Maltese *(pp. 6, 7)*

NL Dutch *(pp. 6, 9)*
NLP Natural Language Processing *(pp. 4, 11, 13)*

PL Polish *(pp. 6, 9)*
PT Portuguese *(pp. 6, 9)*

RNNs Recurrent Neural Networks *(p. 11)*
RO Romanian *(pp. 6, 9)*

SK Slovak *(pp. 6, 9)*
SL Slovene *(pp. 6, 9)*
SV Swedish *(pp. 6, 9)*

UK United Kingdom *(pp. 3, 4, 7)*
USA United States of America *(p. 3)*

VADER Valence Aware Lexicon and sEntiment Reasoner *(p. 13)*

1

Introduction

In politics, words can serve as a key tool to mold public sentiment and support polarizing opinions. Weaponized speeches refer to discourses based on emotional manipulation, divisive rhetoric, and disinformation.

Weaponized speeches, typically associated with social networks or informal situations like interviews, can also emerge in institutionalized and controlled environments such as the European Parliament (EP). Despite this structured setting, political exchanges can escalate into a verbal confrontation, illustrating how words can become weapons in the arsenal of political debate. Such moments reveal the potency of language in politics and the capacity of speech to transcend the boundaries of decorum and ignite significant controversy and publicity. On 2 July 2003, the EP became the stage of a widely reported example of such contentious rhetoric in a formal political context. During Italy's rotating presidency of the Council of the European Union, former Italian Prime Minister, Silvio Berlusconi, was accused by Martin Schulz, a German Member of the European Parliament (MEP) from the social-democrat group, of a conflict of interest because of his media empire. Berlusconi then responded with the comment: " Mr Schultz, I know there is a producer in Italy who is making a film about Nazi concentration camps. I will suggest you for the role of guard. You would be perfect!"[1]. Berlusconi's statements that challenged the decorum of the EP did not stop there, in fact, when the entire chamber showed its disapproval of him, he called all MEPs "tourists of democracy", questioning their commitment to democracy.

Episodes like this, in an institutional setting such as the EP, highlight the potential for weaponized rhetoric and emotional outbursts that can significantly influence the perception of a speech. For this reason, it can be interesting to use sentiment analysis, a tool that can quantitatively measure the emotional tone behind words, in these political discourses. Sentiment analysis provides valuable insights into parliamentary debates and speeches, showing the impact of language in the political sphere. Considering the linguistic diversity of the European Union (EU), with its multitude of official languages

[1]Translation into English from the verbatim report of the EP plenary of 2 July 2003.
Original version in Italian: "Signor Schulz, so che in Italia c'è un produttore che sta montando un film sui campi di concentramento nazisti: la suggerirò per il ruolo di kapò. Lei è perfetto!"

present in the EP discussions, this thesis seeks to explore two questions. First, how is a language interpreted in the context of a political dialogue? And second, does the translation of speeches from one language to another affect the original sentiment? To address these questions, speeches from the EP in 21 official European languages will be examined. The aim is to understand whether there are losses in the translation process that could potentially alter the intended emotional weight of a speech and its consequent political impact.

2

Related Work

Emotions significantly influence the extent to which sentiments become polarized, as Prinz (2021) suggests. Polarization, in essence, is an emotional phenomenon, with Wagner (2021) defining it as the division of society into two distinct groups: those who support a particular political party (partisanship) and those who strongly oppose the rival parties, a concept often referred to as affective polarization (Kim et al., 2023). Historically, especially in the United States of America (USA), polarization was primarily understood as the divergence in political views between Democrats and Republicans (Fiorina et al., 2005). However, it is crucial to distinguish affective polarization from ideologism. Despite their interconnections, these are separate entities both in theory and practice (Iyengar et al., 2019). Kolster et al. (2021) further note that polarization's primary consequence is the tendency of political parties to develop more radical ideas, leading to an increased likelihood of employing provocative and confrontational language in political discourse.

Understanding the concept of polarization prompts an intriguing investigation into the possible escalation of polarization among EU countries in recent times. Significant studies in this area, including those by Jensen et al. (2012), Boxell et al. (2022), Müller and Schnabl (2021) and Volkens et al. (2017), have utilized distinctive datasets and analytical approaches. These comprise analyzing Congressional records and political surveys in the USA to evaluating public posts on social media platforms and scrutinizing political manifestos and electoral results in Europe. Together, these studies provide a comprehensive and diverse outlook on the evolving political landscape, presenting valuable insights into the trends and changes in polarization over time.

Continuing this exploration, a noteworthy contribution is the innovative methodology employed by Caravaca et al. (2022), which centers around the Polarization Index developed by Dalton (2008). This index is very effective for making cross-national comparisons and considers both parties' relevance and positions on a political scale. They calculated polarization utilizing a dataset of public posts from the Facebook pages of 234 political parties across the 27 EU member states and the United Kingdom (UK). This dataset, covering the period from 2019 to 2021, was analyzed in tandem

with the EU parliament elections of May 2019, enabling the researchers to effectively model the polarization within these nations. In their latest update on 15 October 2023, their interactive website dashboard[1], revealed France as the country exhibiting the highest level of polarization, followed closely by Cyprus and Spain. Upon reviewing the temporal dashboards for each country from 1 January 2019, it appears that interesting trends are emerging: Romania, Slovenia, and the UK have seen the largest increases in polarization, with rises of 12.5%, 10%, and 7% respectively. In contrast, Latvia, Germany, and Ireland have witnessed notable decreases in polarization, with respective reductions of 15%, 6.5%, and 5.5%. These results highlight the fluidity of political polarization throughout Europe, providing a valuable understanding of the changing political landscape.

Delving further into the array of methodologies for assessing political polarization, Németh (2023) reviews three Natural Language Processing (NLP) techniques: topic modeling, sentiment analysis, and word embedding methods. Illustrating the application of the first two, Costa (2023) presents an insightful case study centered around Brazil's 2022 presidential elections and its reflection in social media. Tweets from key presidential candidates have been collected and analyzed, using topic modeling to systematically categorize them into thematic groups. Additionally, sentiment analysis has been employed to assess the emotional tones of these tweets. Regarding the utilization of word embedding methods, studies such as Liu et al. (2021) and Rozado and Al-Gharbi (2022) demonstrate the effectiveness of word embedding techniques in analyzing polarization and sentiment associations in different media formats.

Following, it is important to highlight a key concern raised by Chang and Bergen (2023). In their survey they cover over 250 studies, emphasizing the predominant focus on English in the language models, which can lead to biases and limitations in how these models are applied. This point is also supported by the findings of Joshi et al. (2020), who note the disparity in NLP research, with an overemphasis on languages with abundant resources, particularly English. Additionally, Wu and Dredze (2020) have identified challenges confronting multilingual Bidirectional Encoder Representations from Transformers (BERT) models in languages with low resources, where their performance is significantly poorer in contrast to higher-resource languages, and particularly so with monolingual BERT models in such low-resource languages. Confirming this, Armengol-Estapé et al. (2021) conducted tests that demonstrated that multilingual models yield better results for under-resourced languages like Catalan than their monolingual equivalents.

To conclude, Ringe, in *"The Language(s) of Politics: Multilingual Policy-Making in the European Union"* (2022), examines the EU's context and investigates how politicians' use of common foreign languages, as well as their reliance on interpretation and translation, impact political dynamics and decision-making processes. This work provides a crucial

[1] https://eupoliticalbarometer.uc3m.es/dashboard/ideology, consulted on 20 November 2023

perspective on the practical implications of multilingualism in political environments.

3

Data

A fundamental goal of the EU is to promote multilingualism, an essential element of European democratic values. This commitment is reflected in the principle that all languages within the EU are treated equally. It also serves to preserve Europe's diverse linguistic heritage and to encourage language learning. The principle of providing EU citizens with official documents in their own language is upheld by the EP's Directorate-General for Translation, which produces translations of all official documents in the EU's 24 official languages. It enables EU citizens to access official documents in their native language[1].

This thesis starts from the source release of the European Parliament Proceedings Parallel Corpus (EPPPC), developed by Koehn, Philipp (2005), which consists of the verbatim reports of the speeches made in the EP's plenary from 1996 to 2011 in its latest version, the seventh[2], dated 15 May 2012, including reports in 21 European languages. These comprehend Bulgarian (BG), Czech (CS), Danish (DA), Dutch (NL),English (EN), Estonian (ET), Finnish (FI), French (FR), German (DE), Greek (EL), Hungarian (HU), Italian (IT), Latvian (LV), Lithuanian (LT), Polish (PL), Portuguese (PT), Romanian (RO), Slovak (SK), Slovene (SL), Spanish (ES) and Swedish (SV). Croatian, Irish, and Maltese (MT) are the only current European official languages that do not feature in the corpus. All members of the EP have the right to choose and speak in any of the official languages when attending parliamentary sessions.

Furthermore, it is worth mentioning the existence of another corpus of the EP, the Digital Corpus of the European Parliament, developed by Hajlaoui et al. (2014), which includes various types of documents produced from 2001 to 2012. However, it excludes verbatim reports to prevent overlap with the previous corpus. To focus on political speeches in the EP, it has been decided to use exclusively the files from the EPPPC.

The total number of files downloaded from EPPPC is 187,072. It has been decided not to take into account in the following steps 99 verbatim reports in French from 1996, which are not available in any other language, and therefore no comparative analysis

[1] https://european-union.europa.eu/principles-countries-history/languages_en, consulted on 2 November 2023

[2] https://www.statmt.org/europarl/, consulted on 12 April 2023

with other languages is possible. It has been observed that, during the period from 2006 to 2011, the EPPPC does not solely rely on verbatim reports of discussions. Instead, it has files containing descriptions of events such as voting processes, compositions of delegations, and topics for discussion. To enhance the data quality, these files have been omitted from further analysis. Figure 3.1A displays two examples of such files that have been removed for subsequent steps.

Table 3.1 shows that from 1996 to 2005, the EPPPC took a certain meticulousness in considering the totality of the verbatim reports, while Table 3.2 demonstrates that between 2006 and 2011, about three-quarters of the original files had to be removed. Figure 3.2 not only indicates a decrease in the rigour of the EPPPC's file management but also a subsequent overall increase in the number of documents. This increase becomes more pronounced after 2006 and especially from 2007 onwards due to the inclusion of documents in the languages of the new EU Member States that joined after 2004, with the publication of the sixth version[3] of the EPPPC on 4 February 2011.

After this filtering process, the final dataset used for the analysis consists of 56,812 verbatim reports in 21 EU languages. Each of them has a specific structure composed of a document markup (<CHAPTER id>), a speaker markup with identified attributes such as identification number, name, language, and affiliation (<SPEAKER id name language affiliation>), but only the first two are present in all files, and a paragraph markup (<P>).

Regarding language identification, the original language of a speech can be specified in two places: as mentioned, it can be within the speaker tag, but sometimes it can be at the beginning of the text, enclosed in parentheses, as shown in Figure 3.1B. During this phase, along with the removal of 231 rows labeled as MT due to the absence of specific analysis models for the Maltese language, there was also a single row tagged with 'UK' which was processed as EN. Also, it is worth pointing out that the tags attached to a document can vary among its different language versions, particularly when the original language of the speech corresponds to the language of the file. This is demonstrated in Figure 3.1C. A review of the EP's website[4] has shown that texts lacking a language tag are originally in English. Whenever a language tag is missing, English has been assigned as the default original language. In addition, in 801 speeches without language indicators, two-letter acronyms that denote political entities or organizations, as illustrated in Figure 3.1D, have been mistakenly identified as language codes. Consequently, the original language tag of these instances is labeled to represent English.

The final dataset has been created, incorporating all the language variants available for the files. The extracted information includes the file name, speaker ID, and, if it is available, the original language, along with the text in all possible language versions. Regarding the latter, only paragraph markups and language tags at the beginning of the

[3] https://www.statmt.org/europarl/archives.html, consulted on 12 April 2023
[4] https://www.europarl.europa.eu/plenary/en/debates-video.html, consulted on 12 April 2023

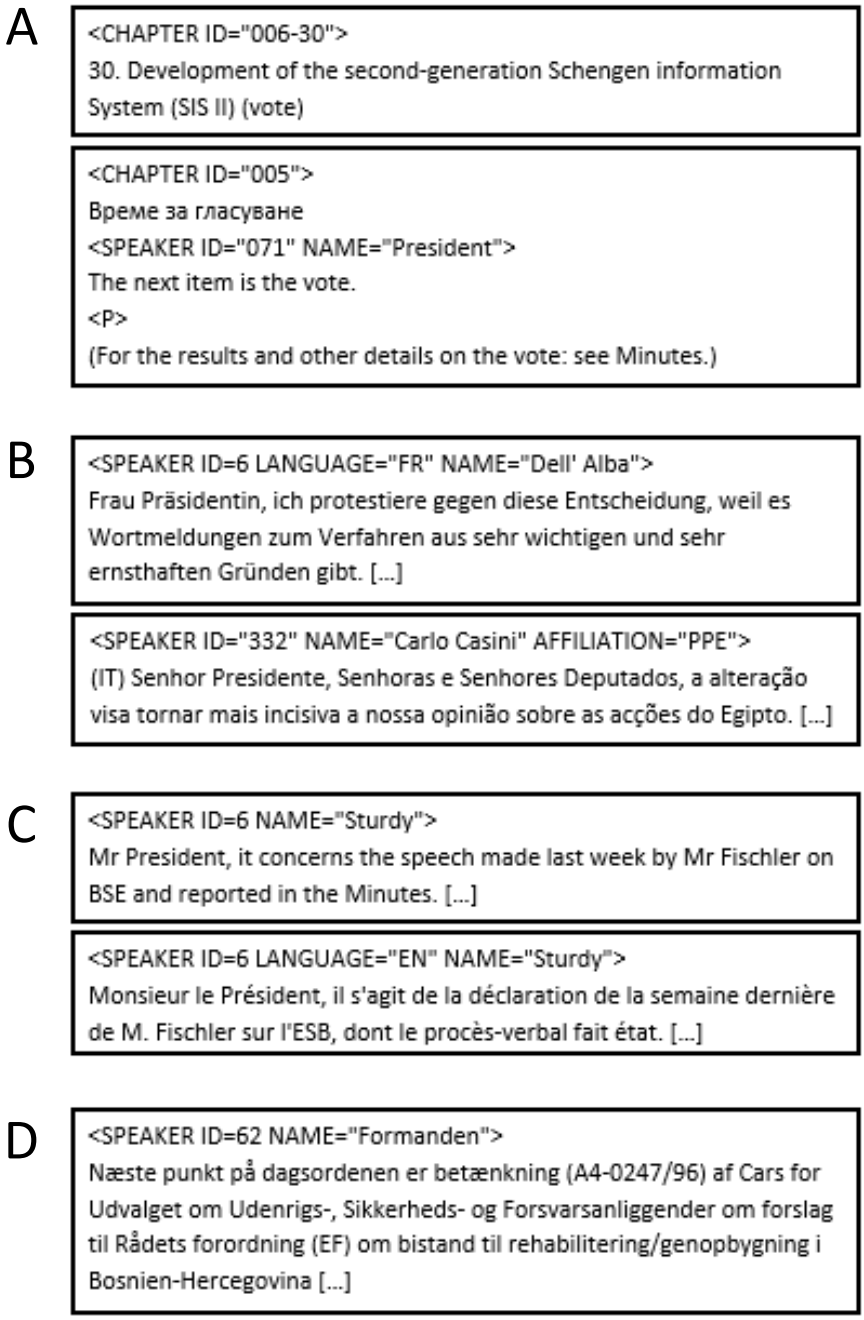

Figure 3.1: Different instances of parts of files from EPPPC. **(A)** Examples of not considered files: en/ep-06-12-14-006-30.txt, English report from 14 December 2006, and bg/ep-08-11-19-005.txt, Bulgarian report from 19 November 2008; **(B)** Examples of language tags: de/ep-00-11-29.txt, German report from 29 November 2000, and pt/ep-10-12-16-012-03.txt, Portuguese report from 16 December 2010; **(C)** Examples of speaker markups with distinct language attributes for different translations of the same speech: en/ep-96-04-15.txt, and fr/ep-96-04-15.txt, respectively English (original for this instance) and French reports from 15 April 1996; **(D)** Example of wrong language tagging, confusing EF (European Foundation) as a proper original language: da/ep-96-07-18.txt, Danish report from 18 July 1996.

LANGUAGE	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
DUTCH	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FRENCH	92%	100%	100%	100%	100%	100%	100%	100%	100%	100%
GERMAN	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
ITALIAN	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
DANISH	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
ENGLISH	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
GREEK	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
PORTUGUESE	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SPANISH	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
FINNISH	-	100%	100%	100%	100%	100%	100%	100%	100%	100%
SWEDISH	-	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 3.1: Percentages of EPPPC Files considered between 1996 and 2005

LANGUAGE	CODE	OFF.	2006	2007	2008	2009	2010	2011
DUTCH	NL	1958	37.44%	30.84%	33.77%	34.88%	28.16%	23.87%
FRENCH	FR	1958	37.64%	31.23%	33.80%	34.54%	28.52%	23.99%
GERMAN	DE	1958	37.74%	31.27%	25.68%	34.91%	29.30%	23.39%
ITALIAN	IT	1958	37.74%	31.27%	33.84%	34.88%	28.93%	23.94%
DANISH	DA	1973	36.93%	31.13%	32.82%	34.88%	28.49%	23.25%
ENGLISH	EN	1973	37.64%	31.15%	33.84%	34.88%	32.80%	26.11%
GREEK	EL	1981	37.44%	31.30%	33.32%	34.88%	28.57%	23.99%
PORTUGUESE	PT	1986	37.54%	31.27%	33.38%	34.50%	29.19%	23.35%
SPANISH	ES	1986	37.54%	30.94%	33.28%	34.76%	29.01%	23.93%
FINNISH	FI	1995	37.74%	30.83%	32.80%	34.88%	28.54%	23.33%
SWEDISH	SV	1995	37.64%	31.10%	33.64%	34.54%	29.64%	23.97%
CZECH	CS	2004	-	17.27%	33.70%	34.88%	29.56%	23.97%
ESTONIAN	ET	2004	-	18.00%	33.50%	34.88%	29.70%	23.60%
HUNGARIAN	HU	2004	-	16.79%	33.80%	34.88%	28.52%	22.31%
LATVIAN	LV	2004	-	18.00%	33.80%	34.88%	28.54%	23.33%
LITHUANIAN	LT	2004	-	16.44%	32.92%	34.88%	28.54%	23.34%
POLISH	PL	2004	0.19%	17.60%	33.72%	34.88%	27.57%	23.67%
SLOVAK	SK	2004	-	16.80%	33.20%	34.88%	28.52%	23.33%
SLOVENE	SL	2004	-	17.80%	33.50%	34.88%	28.52%	23.33%
BULGARIAN	BG	2007	-	-	-	25.30%	27.11%	22.70%
ROMANIAN	RO	2007	-	-	-	24.66%	27.11%	23.34%

Table 3.2: Percentages of EPPPC files considered between 2006 and 2011

text have been removed. To emphasize that, if the speaker's original language used in the speech matches the language of the considered file, then the text in the file represents the original text rather than a translated version. The next step involves merging every file's dataset in different languages based on the file name, speaker ID, and original language to ensure data homogeneity. The final analysis utilizes 3468 unique files, with an annual distribution depicted in Figure 3.3. As previously mentioned, it is observed that the volume of verbatim reports taken into account notably increases starting from the years 2006-2007. The total amount of speeches in the dataset is 202,773.

In conclusion, Table 3.3 presents an example of the merged dataset, showcasing the synthesis of the multilingual files. Further insights into the applied methods will be presented in the forthcoming chapter.

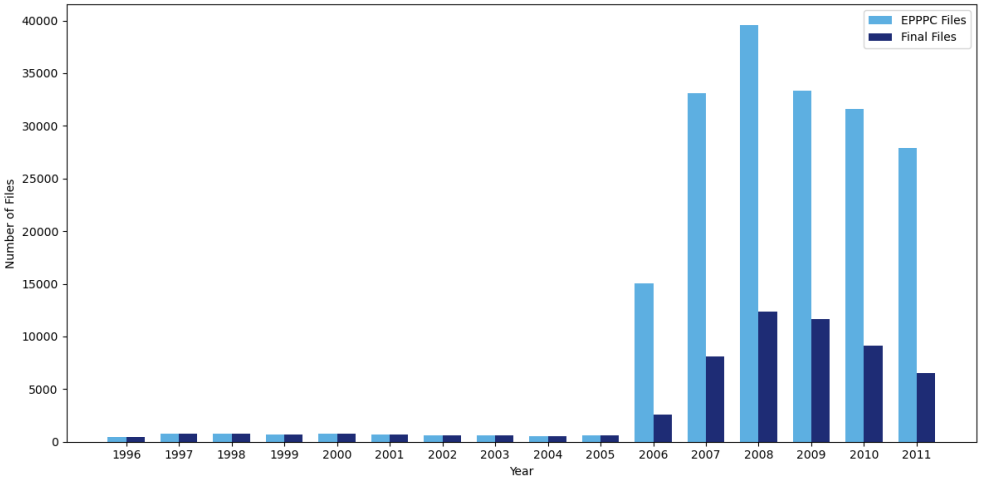

Figure 3.2: Amount of EPPPC files Vs Final files considered per year

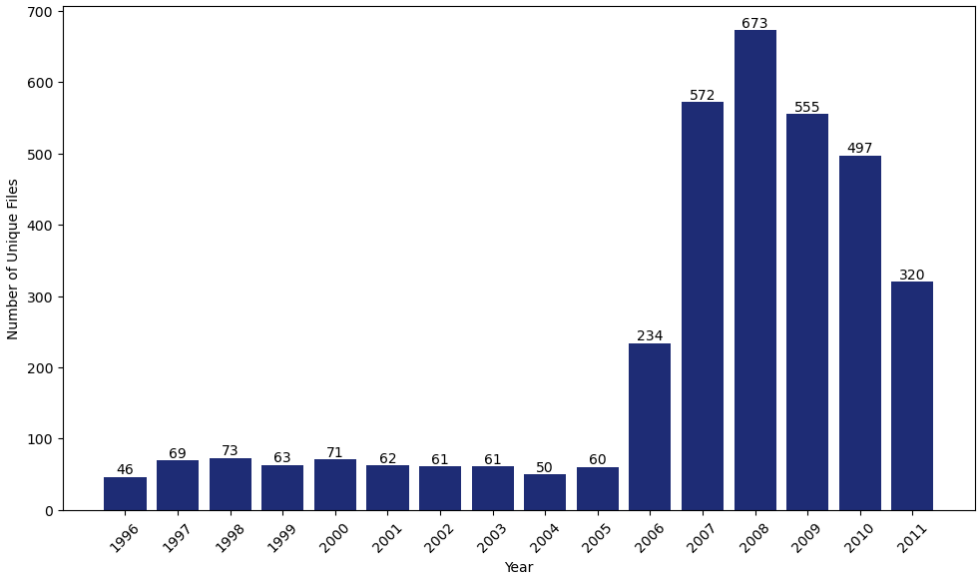

Figure 3.3: Amount of unique Final files considered per year

FILE	SPEAKER ID	ORIGINAL	TEXT_DA	TEXT_DE	TEXT_EL	TEXT_EN	TEXT_ES	TEXT_FR	TEXT_IT	TEXT_NL	TEXT_PT
ep-96-04-15	1	EN	Jeg erklærer Europa-Parlamentets session, der ...	Ich erkläre die am Donnerstag, den 28. März 19...	Κηρύσσω την επανάληψη της συνόδου του Ευρωπαϊκ...	I declare resumed the session of the European ...	Declaro reanudado el período de sesiones del P...	Je déclare reprise la session du Parlement eur...	Dichiaro ripresa la sessione del Parlamento eu...	Ik verklaar de zitting van het Europees Parlem...	Declaro reaberta a sessão do Parlamento Europe...
ep-96-04-15	2	EN	Mine damer og herrer, det er mig en stor glæde...	Liebe Kolleginnen und Kollegen! Im Namen unser...	Αγαπητοί κυρίες και κύριοι συνάδελφοι, εξ ονόμ...	Ladies and gentlemen, on behalf of the House I...	Deseo dar la bienvenida a los miembros de una ...	Chers collègues, je souhaite, au nom du Parlem...	Onorevoli colleghi, a nome della nostra Assemb...	Dames en heren, namens het Parlement verwelkom...	Caros colegas! Em nome do nosso Parlamento saú...
ep-96-04-15	3	EN	Protokollen fra mødet torsdag, den 28. marts 1...	Das Protokoll der Sitzung vom Donnerstag, den ...	Τα συνοπτικά πρακτικά της συνεδρίασης της Πέμπ...	The Minutes of the sitting of Thursday, 28 Mar...	El Acta de la sesión del jueves 28 de marzo de...	Le procès-verbal de la séance du jeudi 28 mars...	Il processo verbale della seduta di giovedì 28...	De Notulen van de vergadering van donderdag 28...	A acta da sessão de 28 de Março de 1996 já foi...
ep-96-04-15	4	NL	Hr. formand, på vegne af mine kolleger i landb...	Herr Präsident, ich möchte Sie im Namen der Ko...	Πρόεδρε, εξ ονόματος των συναδέλφων μου της Επ...	Mr President, on behalf of my fellow-members f...	Señor Presidente, en nombre de los miembros de...	Monsieur le Président, au nom de mes collègues...	Signor Presidente, a nome dei colleghi della c...	Voorzitter, uit naam van de collega's uit de l...	Senhor Presidente, em nome dos meus colegas da...
ep-96-04-15	5	EN	Det må jeg først lige få afklaret, fru Oomen-R...	Das muß ich erst einmal klären, Frau Oomen-Rui...	Αυτό θα πρέπει πρώτα να το ξεκαθαρίσω, κυρία O...	I will have to look into that, Mrs Oomen-Ruijt...	Primero debo aclararlo, señora Oomen-Ruijten. ...	Je dois examiner cela, Madame Oomen-Ruijten. J...	Devo prima informarmi io stesso, onorevole Oom...	Mevrouw Oomen-Ruijten, dat moet ik eerst nagaa...	Tenho de esclarecer o assunto, Senhora Deputad...

Table 3.3: Head of the merged dataset composed by pre-processed files

4

Modeling

After having created the dataframe, it is possible to commence the sentiment analysis, which is the process of extracting emotions from text. Tan et al. (2023) describe a deep learning approach. Wankhade et al. (2022), on the other hand, consider lexicon-based, machine learning, and hybrid approaches as various methods for measuring sentiment.

4.1 Deep Learning Approach

Different model architectures are employed in deep learning techniques for sentiment analysis. In recent literature, Zhang et al. (2018) illustrates the effectiveness of integrating neural network structures, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNNs), and Long Short Term Memory (LSTM), to enhance the interpretative power of sentiment analysis models. Later, Balakrishnan et al. (2022) extend the comparison to include not only the earlier cited neural network models but also other bidirectionally trained models like Bi-directional Long Short Term Memory (Bi-LSTM) and transformer-based like BERT and its variants (RoBERTa and ALBERT).

Acknowledging the capability of transformer models, pioneered by Vaswani et al. (2023), to process context and meaning by tracking relationships in sequential data as the words in a sentence, giving the opportunity to have profound insights into the contextual nuances of a language, it has been decided to use for this thesis BERT, developed by Devlin et al. (2018) at Google.

BERT's success is justified by its double-simultaneous-tasks training: the masked language modeling technique, where it learns by predicting words that are intentionally hidden, and the next sentence prediction, where the model tries to solve a binary classification problem, trying to identify whether the second sentence is effectively the next one or simply random. Given its bidirectional nature, BERT considers context from both sides of a word. BERT's pre-trained language models are suitable to tackle various NLP tasks, achieving state-of-the-art performance.

They are hosted on Hugging Face[1], a collaborative data science platform that

[1] https://huggingface.co/models, consulted on 15 April 2023

deploys open-source models (Wolf et al., 2019). Two varieties of BERT models are available: cased and uncased. The first one keeps the case of the original text, while the uncased version is all lower-case. Jiang et al. (2020) explain that generally, uncased BERT models tend to outperform and provide more consistent results compared to cased models. But as showed by Huang and Cole (2022), in fields with a vast presence of acronyms, such as the scientific with its elements, the BERT cased model achieved the highest precision score. Even though BERT is context-aware, case sensitivity can still be important because capitalized words often have specific meanings or refer to proper nouns. For instance, correctly identifying 'NATO' as the alliance and not 'nato', which means born in Italian, is essential for accurate sentiment analysis.

Given the formal setting of the EP, with numerous proper nouns and acronyms, it has been decided to use an uncased multilingual BERT model processed by NLP Town[2], `'bert-base-multilingual-uncased-sentiment'`[3]. Recognizing that a pre-trained model must be fine-tuned on a task to deliver valuable results (Cheng & Gangaraju, 2023), this particular model has been specifically adjusted based on product reviews in 6 languages: Dutch, English, French, German, Italian, and Spanish. Considering the presence of 21 languages in the corpus, an expansion of the analyzed languages for this model is desired. Thus, the model has been applied to texts in languages from the same Indo-European families as the fine-tuned languages. This means that in addition to the three Germanic and three Latin languages already fine-tuned, Danish and Swedish (Germanic), along with Romanian and Portuguese (Latin), are added to the analysis despite not being individually fine-tuned. In Table A.1 of the Appendix A, it is possible to see a further clarification of the languages considered in each family. The forthcoming results chapter will offer interesting insights when examining how the languages that are not fine-tuned stack up against those that are. Each text has been assigned a sentiment value using the BERT model, which estimates the probability of sentiment classes from very negative (1 star) to very positive (5 stars). These probabilities are then used to compute a weighted average score, which is subsequently normalized to a -1 to 1 range. This process converts categorical sentiment ratings into a continuous scale, enhancing the granularity of sentiment analysis. Only texts with original language tags corresponding to one of the ten languages identified have been included, culminating in a dataset of 178,969 speeches specifically tailored for this model.

4.2 Lexicon-based and Hybrid Approaches

A conventional approach is the lexicon-based method, which was first introduced by Turney (2002) to compute the semantic orientation of sentences in the context of travel reviews. According to Taboada et al. (2011), lexicons consist of an array of

[2] https://nlp.town, consulted on 8 November 2023

[3] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment, consulted on 8 November 2023

Language	Model	Word Vectors
DUTCH	`nl_core_news_lg`	FastText
FRENCH	`fr_core_news_lg`	FastText
GERMAN	`de_core_news_lg`	FastText
ITALIAN	`it_core_news_lg`	FastText
DANISH	`da_core_news_lg`	FastText
ENGLISH	`en_core_web_lg`	FastText
GREEK	`el_core_news_lg`	FastText
SPANISH	`es_core_news_lg`	FastText
PORTUGUESE	`pt_core_news_lg`	FastText
FINNISH	`fi_core_news_lg`	Floret
SWEDISH	`sv_core_news_lg`	Floret
POLISH	`pl_core_news_lg`	FastText
LITHUANIAN	`lt_core_news_lg`	FastText
SLOVENE	`sl_core_news_lg`	Floret
ROMANIAN	`ro_core_news_lg`	FastText

Table 4.1: Used spaCy language models and respective word vectors

tokens, where each token is assigned a specific score indicating whether the language used is neutral, positive, or negative, thereby determining the sentiment conveyed by the text. This method relies on pre-constructed libraries where words are tagged with sentiment values. Textblob[4] and Valence Aware Lexicon and sEntiment Reasoner (VADER) are two of the most popular lexicon-based sentiment analysis libraries in Python. The first determines sentiment scores by averaging the polarity of words within the text (Oyebode & Orji, 2019). Polarity scores range from -1 to 1, indicating the sentiment from negative to positive. Additionally, subjectivity scores from 0 to 1 are provided to indicate the level of objectivity present in the text. On the other hand, VADER, developed by Hutto and Gilbert (2014), detects polarity and emotions specialized mainly in texts from social media characterized by non-formal text, also treating symbols like hashtags and emoticons (Al-Shabi, 2020). For this purpose, the Textblob library has been selected between the two to conduct sentiment analysis since the speeches in the EP are in a formal context.

As the dataset contains texts from 21 different languages, a method capable of performing sentiment analysis across such a wide range of languages is necessary. While Textblob primarily supports English, it also offers extensions for French and German. However, spaCy, a Python-based library developed by Honnibal and Montani (2017), surpasses it for its extensive support of more than 70 languages in various efficient NLP tasks. The `'spacytextblob'` pipeline uses a hybrid approach for measuring sentiment analysis. This integration enables TextBlob to employ its sentiment analysis capabilities on the broad range of linguistic processing features provided by spaCy.

On the starting dataset, it has been decided to apply two models. The first one

[4] https://textblob.readthedocs.io/en/dev/, consulted on 19 April 2023

includes polarity and subjectivity metrics calculated using models for each text's language. However, metrics were not calculated for Bulgarian, Czech, Estonian, Hungarian, Latvian, and Slovak languages due to the unavailability of specific models. After eliminating records labeled with those six languages as their original language, the dataset has been refined to a total of 196,485 speeches.

spaCy's models come pre-trained, machine-learning-based, and are labeled based on the size of word vectors included: `'sm'` for small, `'md'` for medium, `'lg'` for large. For tasks that require a certain depth of linguistic understanding, larger-sized models are generally preferred.[5] For this reason, it has been decided to use only large-sized models, as shown in Table 4.1. This table describes the types of vectors used by each spaCy model. For the majority of languages, FastText vectors[6] are used, which capture a wide range of linguistic nuances. FastText, as introduced by Bojanowski et al. (2017), analyzes words down to their subword components, such as prefixes and suffixes. For languages with complex morphology, spaCy employs Floret vectors [7], which build on FastText's subword analysis by integrating the space-saving Bloom filter technology, a concept introduced by Bloom (1970). They compactly encode this subword information, providing a dense and rich word representation, into spaCy's pipelines, offering compressed vectors up to ten times smaller than traditional ones. Boyd and Wamerdam (2022), in their analysis comparing English with a morphologically rich language such as Korean, observed a significant difference in performance between Floret and FastText vectors. Floret vectors outperformed in the complex Korean language, while in English, the difference was minimal, leading them to continue using FastText for less morphologically rich languages within spaCy. Within the range of languages evaluated, Finnish, Slovene, and Swedish make use of Floret vectors. All models, whether using FastText or Floret, are structured in a 300-dimensional vector space, with FastText models typically having 500,000 keys and unique vectors and Floret models having 200,000 keys, facilitating unified linguistic representation.

The second model is spaCy's `'xx_sent_ud_sm'` transformer-based multilingual to compute both polarity and subjectivity. It is grounded in the Universal Dependencies framework, specifically version 2.8, an international project spearheaded to develop uniform treebank annotations across a wide range of languages (de Marneffe et al., 2021). Notably, this version covers 89 languages, encompassing all 21 languages used in the EPPPC. Consequently, for this model, the dataset retains the entirety of the original, amounting to 202,773 speeches.

These methodologies set the stage for the next results chapter, where the effectiveness and insights drawn from these diverse linguistic analyses will be thoroughly explored and discussed.

[5] https://spacy.io/models, consulted on 19 April 2023
[6] https://github.com/facebookresearch/fastText, consulted on 5 November 2023
[7] https://github.com/explosion/floret, consulted on 5 November 2023

5

Results

This chapter focuses on the comparison of sentiment and subjectivity in texts, analyzing the differences between when a language is the original and when it is the target of translation. In particular, it evaluates how these aspects are captured by the three methods described in the previous chapter.

The graph in Figure 5.1 shows the total number of texts for each language and the proportion of those texts that are in the original language. The length of the bars for the original texts compared to the total number of texts illustrates the extent to which translations are included in the dataset for each language. The texts in the first ten languages have scores comparable to TextBlob's polarity, calculated by the multilingual BERT model, highlighted by the blue line. In addition, texts in the next five languages, including Slovene, have both polarity and subjectivity scores determined by spaCy's monolingual models, represented by the yellow line. Finally, spaCy's multilingual model extends the analysis to all languages and provides metrics for both polarity and subjectivity, indicated by the grey line. The limited availability of models for all languages results in partial analyses for two of the three methods, as they don't use the full range of texts in the corpus. Specifically, the multilingual BERT model excludes 896,067 texts (29.1% of the total), of which 17,717 (9.4% of the total) are in their original languages, while the single-language spaCy models exclude 364,148 texts (11.8% of the total), of which 4,322 (2.2% of the total) are in their original languages.

Table 5.1 introduces an additional level of analysis, showing the pairing of texts from an original language with other translated versions. This information lays the groundwork for deeper exploration in subsequent analyses, enhancing the understanding of how languages interact and influence each other in the multilingual environment of the EP and exploring the broader implications of the reach of the methods.

The first analysis is a temporal study, looking at fluctuations in sentiment and subjectivity from 1996 to 2011, using different methods for each language where applicable. Polarity can be measured by all three methods, while subjectivity is limited to the two spaCy models. Furthermore, not all languages are available for each method. Therefore, the analysis has been divided into three distinct groups based on the available

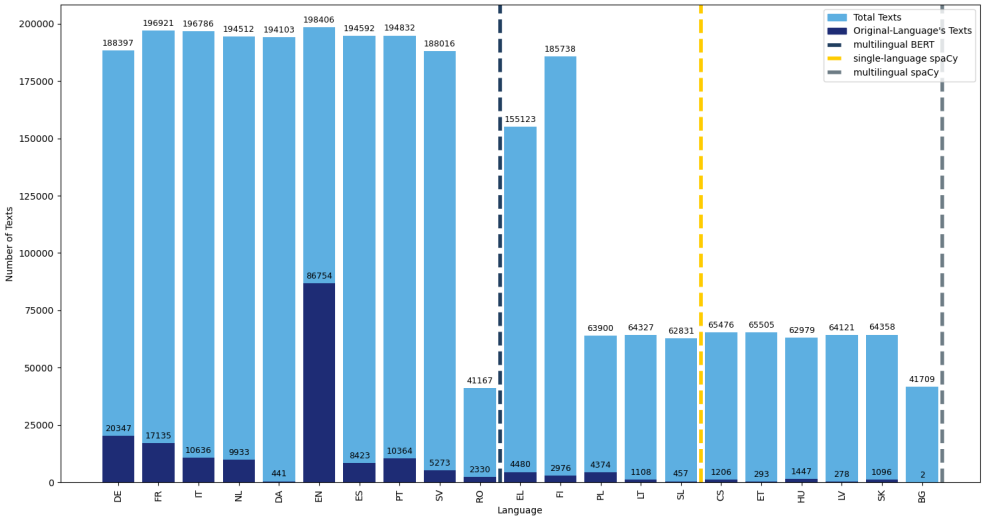

Figure 5.1: Amount of total texts available in a language, frequency of speeches in that language as original versions, and corresponding used sentiment analysis methodologies

Original Language \ Translated Language	DE	FR	IT	NL	DA	EN	ES	PT	SV	RO	EL	FI	PL	LT	SL	CS	ET	HU	LV	SK	BG
DE	20.3	21.0	21.0	20.8	20.9	21.2	20.8	20.9	20.2	4.2	16.9	20.0	6.4	6.5	6.3	6.6	6.6	6.3	6.4	6.5	4.3
FR	16.3	17.1	17.1	16.9	16.9	17.3	16.9	17.0	16.4	3.7	13.9	16.1	5.8	5.9	5.7	6.0	6.0	5.8	5.8	5.9	3.7
IT	10.2	10.6	10.6	10.5	10.5	10.8	10.5	10.5	10.2	3.1	8.6	10.1	4.1	4.1	4.0	4.2	4.2	3.9	4.1	4.1	3.2
NL	9.6	10.0	10.0	9.9	9.9	10.0	9.9	9.9	9.4	1.2	7.8	9.3	2.1	2.1	2.0	2.1	2.1	2.0	2.1	2.1	1.2
DA	0.4	0.5	0.5	0.5	0.4	0.5	0.5	0.5	0.5	0.3	0.5	0.5	0.4	0.5	0.4	0.5	0.5	0.4	0.4	0.4	0.3
EN	83.6	86.8	86.7	85.7	85.4	86.8	85.4	85.6	81.6	12.4	64.6	80.9	19.9	20.1	19.6	20.3	20.3	19.6	20.0	20.1	12.6
ES	8.2	8.5	8.5	8.4	8.4	8.5	8.4	8.5	8.1	1.3	6.5	8.0	2.1	2.2	2.1	2.2	2.2	2.1	2.1	2.2	1.3
PT	9.8	10.4	10.5	10.2	10.3	10.7	10.3	10.4	10.2	4.0	8.7	10.1	5.3	5.3	5.3	5.5	5.5	5.2	5.3	5.4	4.0
SV	5.2	5.6	5.5	5.5	5.5	5.6	5.5	5.5	5.3	0.8	4.1	5.2	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	0.8
RO	2.9	3.1	3.2	3.1	3.1	3.3	3.2	3.1	3.2	2.3	3.1	3.1	3.1	3.2	3.1	3.2	3.2	3.1	3.1	3.1	2.4
EL	5.5	5.8	5.7	5.7	5.7	5.8	5.7	5.7	5.6	1.1	4.5	5.5	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.1
FI	3.0	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.0	0.5	2.4	3.0	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.5
PL	5.7	6.2	6.2	6.1	6.1	6.3	6.2	6.2	6.2	2.4	5.8	6.0	4.4	4.4	4.3	4.5	4.5	4.3	4.4	4.4	2.4
LT	1.4	1.5	1.5	1.4	1.4	1.5	1.4	1.4	1.4	0.9	1.4	1.4	1.1	1.1	1.0	1.1	1.1	1.0	1.1	1.1	0.9
SL	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.2	0.6	0.6	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.2
CS	1.6	1.7	1.7	1.6	1.6	1.7	1.7	1.7	1.7	0.8	1.6	1.6	1.2	1.2	1.1	1.2	1.2	1.2	1.2	1.2	0.8
ET	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.2	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2
HU	1.9	2.0	2.0	1.9	1.9	2.0	2.0	2.0	2.0	1.0	1.9	1.9	1.5	1.5	1.4	1.5	1.5	1.4	1.5	1.5	1.0
LV	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.2	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.2
SK	1.3	1.5	1.5	1.4	1.4	1.5	1.4	1.5	1.5	0.7	1.4	1.4	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	0.7
BG	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.0

Table 5.1: Quantity in thousands of texts for each original and translated language pair

methods and the languages that can be compared within each method.

Looking at the polarity for the first ten languages in Figure 5.2, it is worth noting that spaCy's models show remarkably identical behaviour over time for both the original and translated versions. This uniformity results in a generally neutral sentiment across all languages, except for English and Romanian, where a slight positive trend is observed. In contrast, while the spaCy and BERT models show similar patterns in Spanish, BERT shows a more varied polarity in other languages, alternating between positive, negative, or mixed sentiments throughout the timeline. In the sentiment analysis, German, Italian, and Danish show a positive trend in the BERT model. It is noteworthy that the upward shift for Danish starts in 2006, when it appears as an original language for the first time.

French and Portuguese, on the other hand, show slightly negative sentiments. Dutch stands out in the BERT analysis as being significantly negative, with a score of around -0.3 when it is the source language and around -0.2 when it is the target language. This is particularly noteworthy for Dutch, as the BERT model used in sentiment analysis is fine-tuned for this language. English and Romanian, which are slightly positive in spaCy's results, are neutral in BERT's analysis for the original language, with Romanian translating slightly negatively. Swedish uniquely shows a wide range of sentiment, fluctuating between -0.2 and 0.2 from 1997 to 2011. Considering subjectivity for the same languages, both German and Dutch consistently show lower levels, around 0.1 to 0.2, in both the original and translated forms. Other languages typically range between 0.2 and 0.4. Notably, French and English maintain a consistent subjectivity of around 0.4 for both their original and translated versions, positioning them at the higher end of the spectrum. In 2006, there is a notable increase in the subjectivity of Danish in its original version, where it reaches a high of 0.6. This represents a significant divergence from its translated counterpart, underscoring the largest observed gap in subjectivity scores between the original and translated versions of any language in the dataset.

In Figure 5.3, the analysis is extended to the remaining languages. The first five are analyzed using both spaCy models, while the last six are analyzed using only the multilingual version. Across all available forms for each language, the behaviour for both polarity and subjectivity is generally close to neutral, reflecting the uniformity seen earlier. A notable exception is Slovene, which deviates from this pattern. In its original form, Slovene exhibits distinct scores, with both polarity and subjectivity around 0.2 in 2007.

Understanding the numerical values in this analysis is essential for an accurate interpretation of the results. In the case of polarity, which ranges from -1 to 1, a change of 0.2 represents a significant shift in sentiment of 10%. This understanding is particularly important when evaluating languages such as Dutch, where the observed scores for both the original and translated versions indicate a deviation of approximately 10 to 15% towards negativity compared to a neutral baseline. In subjectivity analysis, where the scale is from 0 to 1, the metric for a 10% change is represented by a smaller magnitude of 0.1. This scaling difference is crucial to consider, especially in the context of pronounced variations such as the spike observed in Danish texts in 2006. In this particular case, where the subjectivity score is particularly high at 0.6, this means that the original Danish texts in 2006 are, on average, more subjective than objective.

The next phase of analysis in this chapter deals with a comparison between the original and translated polarity and subjectivity scores for each language using the multilingual spaCy model. This particular model is used because of its completeness in terms of the languages covered and because, as previously demonstrated, the two spaCy methodologies produce essentially identical results for most languages. This comparison aims to determine whether there are consistent patterns or notable differences in average sentiment scores across language families and how translation

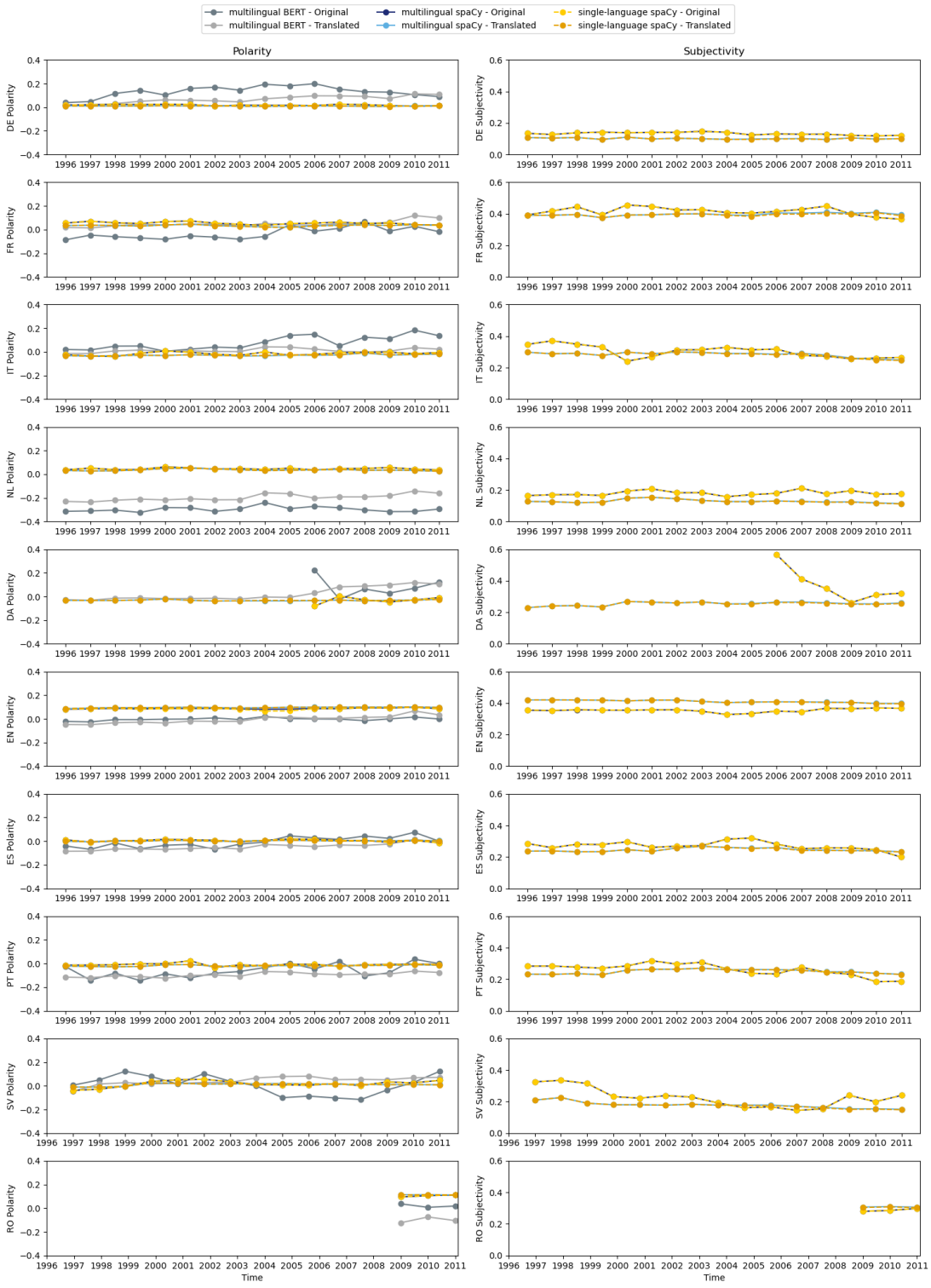

Figure 5.2: Behaviour of polarity and subjectivity between 1996 and 2011 for the ten languages considered by the three methodologies

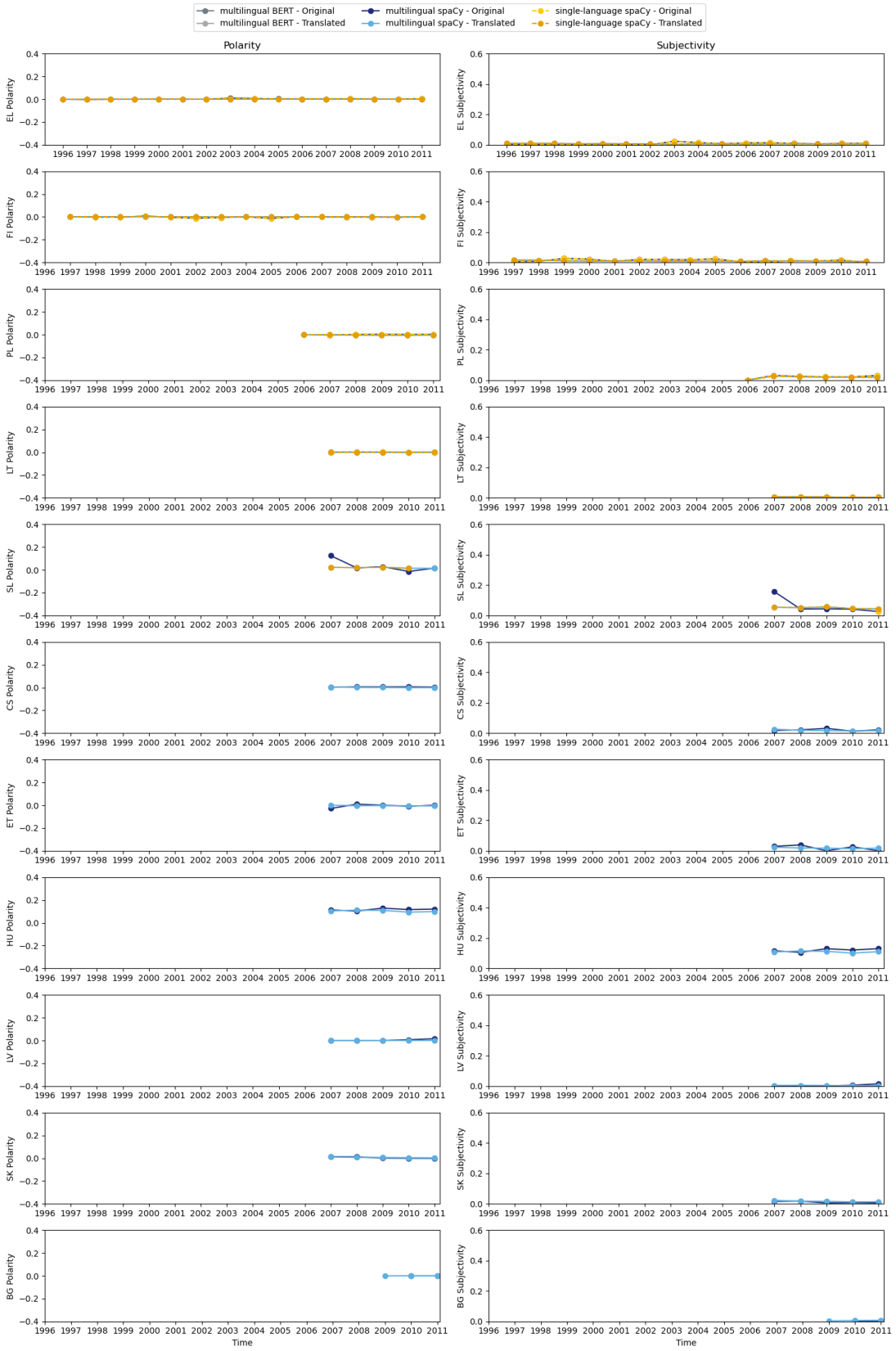

Figure 5.3: Behaviour of polarity and subjectivity between 1996 and 2011 for the other eleven languages considered by the two spaCy's methodologies

affects these scores. The analysis focuses on the position of the language data points relative to the diagonal of the Cartesian plane, which indicates an equal sentiment score in both the original and translated texts. If languages appear to the left of the diagonal, this suggests that translation increases their polarity or subjectivity. On the other hand, placement to the right suggests that the sentiment is more pronounced in its original form than in its translated form.

In the Cartesian plane of Figure 5.4A, Romanian, English, and Hungarian show robust polarity values, hovering around 0.10 to 0.12 for both original and translated versions, with the first two languages well to the left of the diagonal, indicating an increase in polarity with translation. Dutch and French are presented with polarity values close to 0.04 and 0.05 for translated and original texts, respectively, indicating a more moderate polarity intensity. A significant concentration of languages, especially Slavic, Baltic, and Hellenic, converge around the neutral centre of the Cartesian plane. This indicates a general tendency towards neutrality in sentiment for these languages, whether in their original form or in translation. Meanwhile, other language families are more evenly distributed along the diagonal, suggesting varying degrees of polarity. Polish and Estonian register only a slight negativity, essentially hovering near a neutral stance. Danish, Italian, and Portuguese have a distinctly negative polarity, but their intensity doesn't match the positive levels observed in Romanian, English, and Hungarian.

Following the analysis of the polarity scores, the focus in Figure 5.4B shifts to subjectivity. Similar to the polarity results, English, Romanian, and this time Portuguese, are placed to the left of the diagonal. This placement suggests an increase in subjectivity in their translated forms. Notably, some Slavic languages also fall on the left side of the diagonal, although they are closer to a neutral position, indicating a less pronounced increase in subjectivity. In contrast to the polarity analysis, the subjectivity scores show a more coherent pattern within the language families. The Germanic and Latin language groups, except English and French, show significantly higher levels of subjectivity. Specifically, four out of the five Latin languages, following the lead of French, which has the highest subjectivity as an original language, register around 0.2 and 0.3 for original and translated scores. Similarly, the remaining Germanic languages, following the trend set by English, which has the highest subjectivity in its original form, have subjectivity scores between 0.15 and 0.3 for both the original and translated versions. In addition, Hungarian, which stands out as a unique case beyond the Germanic and Latin language families, falls within this range. In contrast to these groups, most of the other languages, especially those from the Slavic, Baltic, and Hellenic families, tend to have neutral subjectivity scores, reflecting a similar position to that observed in their polarity analysis.

Following, for each language, the standard error of the mean changes in polarity and subjectivity due to translation from the original language is calculated. A positive difference indicates that translation increases sentiment, while a negative difference

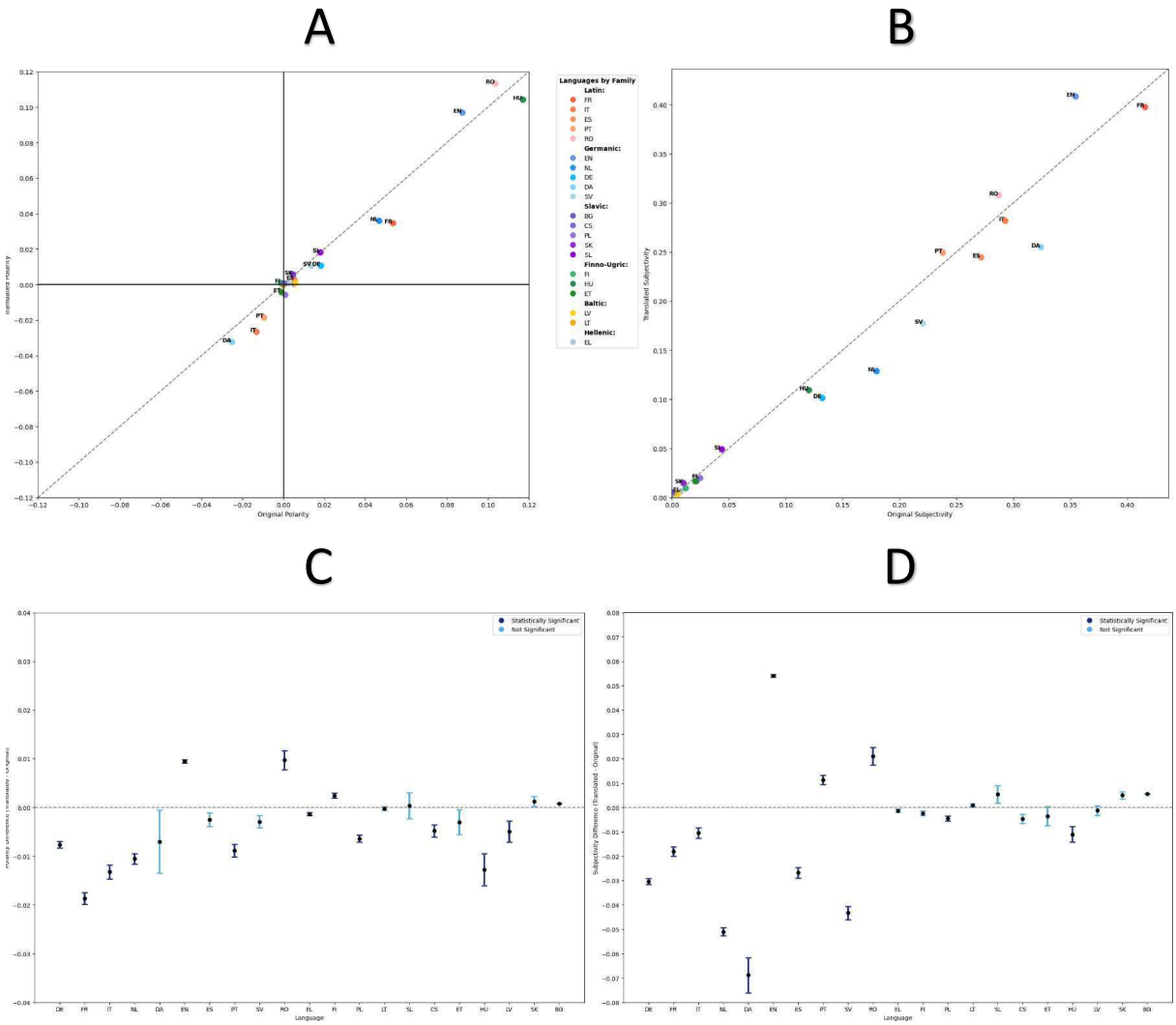

Figure 5.4: Comparison and statistical significance of polarity and subjectivity scores and shifts across language translation**(A)** Comparison of polarity scores for original versus translated texts across language families; **(B)** Comparison of polarity scores for original versus translated texts across language families; **(C)** Statistical significance of polarity shifts in text translation across languages; **(D)** Statistical significance of subjectivity shifts in text translation across languages

suggests a decrease. A two-tailed t-test is then utilized to determine the statistical significance of these shifts, thereby examining how the translation process influences sentiment metrics in the considered languages.

Figure 5.4C and Figure 5.4D show the effects on polarity and subjectivity, respectively, across different languages. A key observation is the relatively narrow range of average polarity shifts between languages, from -0.02 (French) to 0.01 (English and Romanian), in contrast to subjectivity, which varies more widely, from -0.07 (Danish) to over 0.05 (English). Notably, Danish has the largest range of standard errors in both categories. In the t-test for polarity and subjectivity across languages, it is noted

that shifts near zero usually have p-values above 0.05, indicating a lack of statistical significance. This trend is consistent across all languages, except for Greek, where even minimal changes in polarity show statistical significance.

In a new analysis, attention is turned to language pairs to assess the impact of translation on sentiment scores. The focus is on correlating polarity scores between language pairs using spaCy's multilingual model, considering all 420 possible combinations from 21 languages. Then, for a comprehensive view, the polarity scores from spaCy's multilingual model are compared with those from BERT's multilingual model. In addition, the study uses only spaCy's multilingual model to measure subjectivity shifts during translation, providing insights into how sentiment, both polarity and subjectivity, is affected in the translation process.

Figure 5.5A shows the correlation between polarity scores from spaCy's multilingual model in different language pairs, where one language is the original and the other is the translated version. It can be seen that the majority of language pairs do not follow a uniform trend, but there are some interesting cases. For example, English, as the dominant language in the dataset with almost half of the original texts, generally shows negative or marginally neutral correlations with most languages, with notable exceptions being moderately positive correlations for translations into Dutch and Portuguese. However, a notable exception to this trend is the translation from English into Bulgarian, which stands out with the highest positive correlation of 0.85, indicating exceptionally effective sentiment preservation in this language pair. In addition, Swedish and Romanian texts translated into Latvian show high positive correlations of 0.84 and 0.76, respectively. In stark contrast, translations from Bulgarian to Czech and Italian to Estonian are characterized by strong negative correlations of -0.92 and -0.9, respectively, indicating a potential inversion of sentiment in these translations. These strong negative correlations indicate a sentiment shift where high sentiment scores in the original language are inversely related to those in the translated language. Such strong negative correlations indicate potential challenges in maintaining sentiment accuracy during translation for these language pairs. Figure 5.5B shows an analysis of sentiment correlations across 45 different language pairs, covering a set of 10 languages and using BERT's model. The predominant pattern observed in this analysis suggests a range of neutral to slightly positive correlations, as in the case of texts originally in Italian and Dutch. A neutral correlation indicates a lack of consistent alignment between the sentiment scores of the original and translated texts. Moving on to the analysis of subjectivity shifts performed with the multilingual spaCy model in Figure 5.5C, a similar pattern of neutral correlations is evident. However, a notable deviation from this trend is seen in the translations of Latvian texts into various Slavic languages, as well as into Finnish and Estonian, where moderate to high positive correlations are recorded. These instances of higher positive correlations suggest a more effective preservation of subjectivity in these specific language translations and stand out against the broader background of neutral correlations observed in other language pairs.

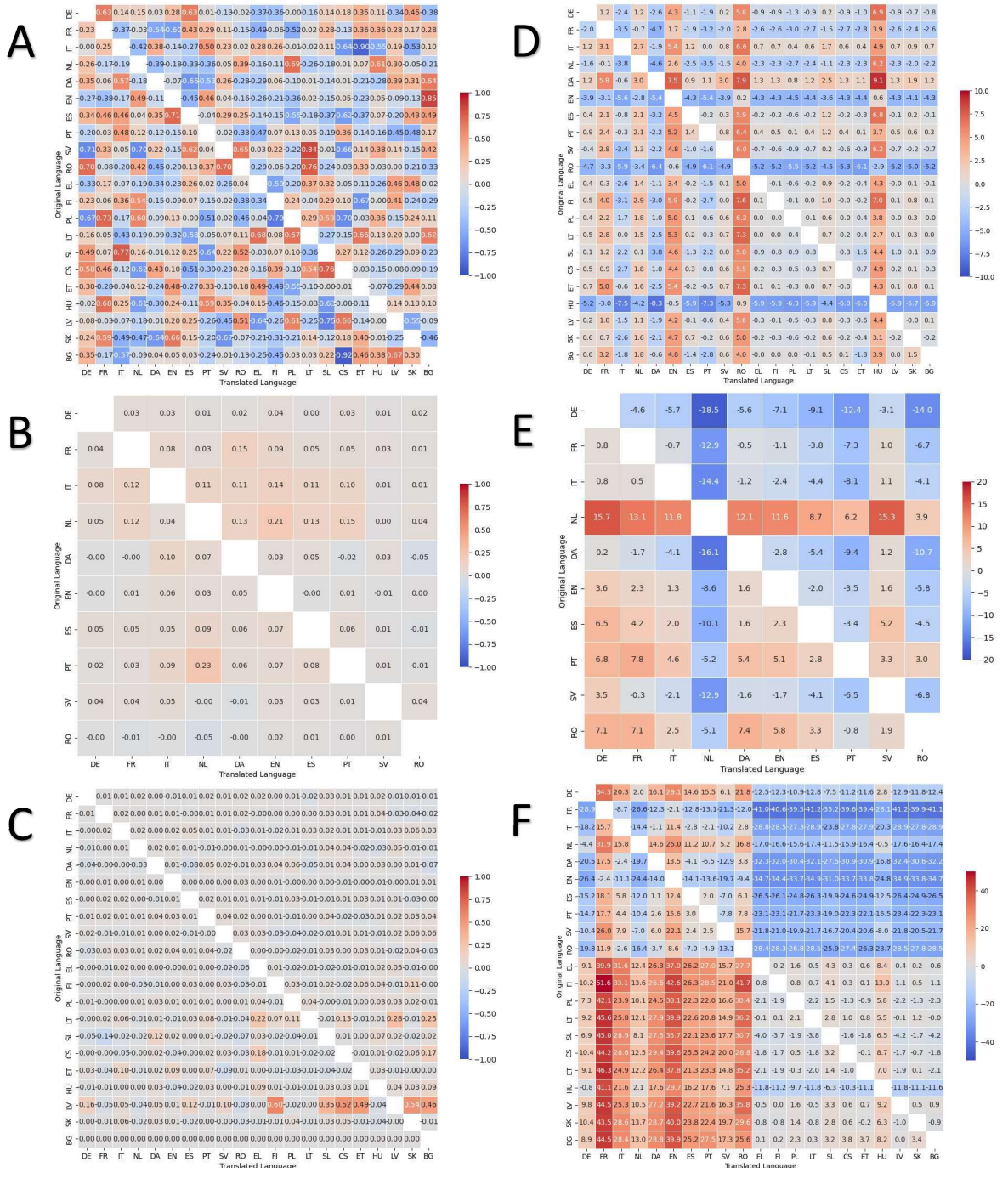

Figure 5.5: Correlation and shifts in polarity and subjectivity across language translations **(A)** Correlation of polarity scores across language pairs using spaCy's multilingual model; **(B)** Correlation of polarity with BERT's multilingual model; **(C)** Correlation of subjectivity scores using spaCy's model; **(D)** Polarity shifts in language translation using spaCy's model; **(E)** Polarity shifts in language translation using BERT's model; **(F)** Subjectivity alterations during language translation with spaCy's model.

In the final section of this discussion of results, the focus shifts to the essence of this thesis: the quantification of sentiment shifts in translation. Using spaCy's multilingual model, shown in Figure 5.5D, several key observations emerge. In particular, English, Romanian, and Hungarian stand out as the languages with the most positive sentiment. This is illustrated by the marked increase in positive polarity of 9.1% when a text in Danish is translated into Hungarian. This phenomenon suggests that the inherent positivity of these languages influences the translated sentiment to such an extent that it is either amplified or diminished whether they are the original or the translated version, as seen in the 8.3% decrease in sentiment when Hungarian is translated back into Danish. A similar pattern can be seen in most language pairs involving these three languages, highlighting a consistent trend in sentiment fluctuation during translation. In contrast to the previously mentioned languages, Italian, Danish, and Portuguese show a more negative sentiment bias. Of these, Danish stands out as the most negative, while Hungarian is at the opposite end of the spectrum as the most positive. This clear contrast between Danish and Hungarian polarities results in the most significant polarity shifts observed among the available language pairings, underscoring the impact of inherent linguistic sentiment on translation. Comparing the sentiment analysis between the BERT model and spaCy's reveals notable linguistic shifts, as shown in Figure 5.5E. Dutch emerges as the most negatively biased language within BERT's framework, a deviation from its slightly positive sentiment in spaCy's model. Portuguese, Romanian, Danish, and Swedish are the four languages that have been considered, even though the BERT model used is not specifically tuned for them in sentiment analysis. Notably, Portuguese is the only language that consistently retains its negative sentiment bias across models. The most striking sentiment shift occurs in translations between German and Dutch. Texts originally in German lose a remarkable 18.5% of their polarity when translated into Dutch, while the reverse translation results in an almost 16% increase in positive sentiment for German versions of Dutch texts. This inversion, where the negative shift exceeds the positive, suggests a complex interplay between language-specific sentiment biases and the nuances of the translation process. The final segment of the analysis, shown in Figure 5.5F, reveals trends in changes in subjectivity during translation. Translations from Hellenic, Finno-Ugric, Baltic, and Slavic languages into Germanic or Latin counterparts typically show an average increase in subjectivity of around 30%. This means that non-Germanic and non-Latin languages are more objective than Latin and Germanic languages. In particular, translations into French show a pronounced average increase in subjectivity of more than 40%, highlighting its increased subjectivity within the spaCy multilingual model, as previously shown in Figure 5.4B. An exceptional case is observed with Finnish texts translated into French, where the increase in subjectivity exceeds 50%. Conversely, within the spectrum of Latin and Germanic languages, German shows the smallest increase in subjectivity, which means it is the most objective. Uniquely, translations from Hungarian into languages outside the Germanic and Latin families

tend to become more objective, with an average decrease in subjectivity of 10%. For the rest of the language pairs analyzed, a predominantly neutral shift in subjectivity is observed.

To properly understand the outcomes that have been observed earlier, it is important to take into consideration that there are two specialized techniques in the EP, "retour" and "relay", that are crucial for facilitating communication with lesser-used languages. The retour method requires interpreters to work from a foreign language into their mother tongue and translate from their second language into their mother tongue. This technique was first introduced in 1995 to meet the challenges of Finnish, which had few non-native speakers who could translate into other languages. Relay translation is more complex and involves translating first into a pivot language before reaching the final target language[1]. The translation processes of return and relay can introduce several layers of bias. As the translation progresses from source to target, each stage can subtly alter the emotions and perspectives conveyed. The translator's choices can affect the fidelity of the translation to the original polarity and subjectivity, leading to variations in the emotional tone or expressiveness of the text. The polarity and subjectivity identified in the source language may shift when expressed in the target language.

This chapter provides a thorough analysis of text frequencies across various language pairs in the extensive corpus of the EP. The chapter's progression moves from a broad temporal analysis to more specific investigations. Initial studies trace shifts in polarity and subjectivity from 1996 to 2011. The findings indicate that spaCy's models predominantly exhibit neutral sentiment over time, whereas subjectivity displays greater variations. BERT, on the other hand, reveals a more diverse range of emotional expression in the polarity of texts across the studied languages. It has been observed that spaCy's monolingual models behave similarly to its multilingual version. Given this consistency and the wider language range of the multilingual models, subsequent analyses concentrate solely on two multilingual models: spaCy and BERT.

The final part of the research focused on how sentiment changes through translation across language families, analyzing polarity and subjectivity to understand sentiment preservation or alteration. For instance, English and Romanian, following spaCy's multilingual model results, show a positive polarity that indicates an optimistic tone in their political discourse. When analyzing polarization trends, it was concluded that there is emotional expression consistency when translating from languages of the same family to others. An example is the observed increase in subjectivity when translating from Baltic, Finno-Ugric, Hellenic, and Slavic languages to their Germanic or Latin counterparts.

[1] https://www.europarl.europa.eu/RegData/presse/pr_focus/2006/EN/03A-DV-PRESSE_FCS(2006)04-03(06935)_EN.pdf, consulted on 15 November 2023

6

Conclusions

In this concluding chapter, the focus is on addressing two primary questions this thesis set out to answer: firstly, understanding how languages are perceived in political discourses, and secondly, evaluating the degree to which the translation process contributes to a 'loss in translation' of the initial sentiment.

When exploring the first question, the analysis concentrated on the polarity of varying languages in a political context. By utilizing spaCy's multilingual model, clear sentiment biases emerged. English and Romanian displayed a consistent positive polarity that reflects an optimistic tone in their political rhetoric. On the contrary, the analysis has exposed that languages such as Danish, Italian, and remarkably Portuguese, as noticed in both spaCy and BERT models, displayed a tendency towards negative polarity. This trend indicates a more reserved or controlled tone in their political discourse. A noteworthy discovery is the significant rise in bias in translations into French, underscoring how it amplifies subjective interpretations in the translation process, implying that French political dialogue may be the most emotional language in the EU. Additionally, research has revealed that Baltic, Finno-Ugric, Hellenic, and Slavic languages exhibit more objectivity compared to Latin and Germanic languages. These variations in emotional expression across languages highlight the diverse nuances present in political communication.

Regarding the impact of translation on original sentiment, significant polarity shifts can be observed across various language pairs in the spaCy model analysis of 420 combinations. It is particularly evident that translations from Bulgarian to Czech and from Italian to Estonian show significant negative correlations, indicating a potential reversal or 'loss' of sentiment during translation. To highlight the common increase in subjectivity when translating from Baltic, Finno-Ugric, Hellenic, and Slavic languages into Germanic or Latin counterparts, whereas Hungarian translations from non-Germanic and non-Latin languages tend to be more objective. These findings demonstrate the complex process of retaining sentiment in translation and accentuate the difficulties of upholding the authentic emotional resonance and political subtleties of speeches in the multilingual environment of the EP.

This study has a major limitation due to the limited number of 10 languages analyzed by `'bert-base-multilingual-uncased-sentiment'`, of which only 6 were fine-tuned for sentiment analysis. The decision to incorporate the remaining 4 languages, chosen for their linguistic similarity to the fine-tuned languages, has not led to any anomalous results. This result implies that selecting languages from the same linguistic families as the fine-tuned models was a wise decision. Nevertheless, it does raise the question of whether including languages that are not linguistically similar to the fine-tuned ones might have led to different results.

Future studies should aim to extend the language coverage by fine-tuning BERT models for all 21 languages present in the corpus. By doing so, it is possible to perform sentiment analysis across a diverse and complete linguistic spectrum. The inclusion of comparative analysis utilizing these recently fine-tuned monolingual BERT models, in conjunction with the existing single-language spaCy models, would provide a richer understanding of polarity shifts in multilingual political discourses.

Considering future research possibilities, the work of Zúquete et al. (2023) is particularly notable. They created a Moral Foundations Dictionary tailored to the Portuguese language, which they then used to analyze and measure the moral aspects of debates in the Portuguese Parliament. Similarly, Geller et al. (2023) demonstrates the use of a speech toxicity classification model to analyze Twitter dynamics. These methods could be valuable models for research in the context of the EP. Adapting such methodologies to the multilingual environment of the EP could significantly enhance the understanding of both moral and toxic elements in political discourse. This approach would enrich the analysis of polarity and subjectivity, aligning with the models utilized in this thesis.

Bibliography

Prinz, J. (2021). Emotion and political polarization. In S. F. Ana & G. da Silva (Eds.). Springer International Publishing. https://doi.org/10.1007/978-3-030-56021-8_1 (cit. on p. 3).

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, *69*, 102199. https://doi.org/https://doi.org/10.1016/j.electstud.2020.102199 (cit. on p. 3).

Kim, Y., Hwang, H., & Kim, Y. (2023). The "us vs. them" mentality: The role of affective polarization in deepening the partisan divide in media bias perception. *Journalism Mass Communication Quarterly*, *0*. https://doi.org/10.1177/10776990231211804 (cit. on p. 3).

Fiorina, M. P., Abrams, S. J., & Pope, J. C. (2005). *Culture war? the myth of a polarized america*. (Cit. on p. 3).

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, *22*, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034 (cit. on p. 3).

Kolster, C., Wittrich, H., & Mauracher, M. (2021). Is european politics polarizing? *Open European Dialogue* (cit. on p. 3).

Jensen, J., Kaplan, E., Naidu, S., & Wilse-Samson, L. (2012). Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. *Brookings Papers on Economic Activity*, *43*(2 (Fall)), 1–81. https://ideas.repec.org/a/bin/bpeajo/v43y2012i2012-02p1-81.html (cit. on p. 3).

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2022). Cross-Country Trends in Affective Polarization. *The Review of Economics and Statistics*, 1–60. https://doi.org/10.1162/rest_a_01160 (cit. on p. 3).

Müller, S., & Schnabl, G. (2021). A Database and Index for Political Polarization in the EU. *Economics Bulletin*, *41*(4), 2232–2248. https://ideas.repec.org/a/ebl/ecbull/eb-21-01168.html (cit. on p. 3).

Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., Weßels, B., & für Sozialforschung (WZB), W. B. (2017). Manifesto project dataset. https://doi.org/10.25522/manifesto.mpds.2017a (cit. on p. 3).

Caravaca, F., González-Cabañas, J., Cuevas, & Cuevas, R. (2022). Estimating ideology and polarization in european countries using facebook data. *EPJ Data Science*, *11*, 56. https://doi.org/10.1140/epjds/s13688-022-00367-1 (cit. on p. 3).

Dalton, R. J. (2008). The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative Political Studies*, *41*, 899–920. https://doi.org/10.1177/0010414008315860 (cit. on p. 3).

Németh, R. (2023). A scoping review on the use of natural language processing in research on political polarization: Trends and research prospects. *Journal of Computational Social Science*, *6*, 289–313. https://doi.org/10.1007/s42001-022-00196-2 (cit. on p. 4).

Costa, F. S. (2023). Analyzing polarization on social media: A case study of the 2022 Brazil presidential election. http://hdl.handle.net/10362/152308 (cit. on p. 4).

Liu, R., Wang, L., Jia, C., & Vosoughi, S. (2021). Political depolarization of news articles using attribute-aware word embeddings. http://arxiv.org/abs/2101.01391 (cit. on p. 4).

Rozado, D., & Al-Gharbi, M. (2022). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science*, *5*. https://doi.org/10.1007/s42001-021-00130-y (cit. on p. 4).

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. http://arxiv.org/abs/2303.11504 (cit. on p. 4).

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world (D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault, Eds.), 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560 (cit. on p. 4).

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? (S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, & H. Hajishirzi, Eds.), 120–130. https://doi.org/10.18653/v1/2020.repl4nlp-1.16 (cit. on p. 4).

Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-Oller, C., Gonzalez-Agirre, A., Melero, M., & Villegas, M. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan (C. Zong, F. Xia, W. Li, & R. Navigli, Eds.), 4933–4946. https://doi.org/10.18653/v1/2021.findings-acl.437 (cit. on p. 4).

Ringe, N. (2022). *The language(s) of politics: Multilingual policy-making in the European Union*. Ann Arbor: University of Michigan Press. https://doi.org/https://doi.org/10.1353/book.99520 (cit. on p. 4).

Koehn, Philipp. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation, 79–86. {https://aclanthology.org/2005.mtsummit-papers.11} (cit. on p. 6).

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., & Varga, D. Dcep - digital corpus of the european parliament. In: 2014-05 (cit. on p. 6).

Tan, K., Lee, C.-P., & Lim, K. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, *13*, 4550. https://doi.org/10.3390/app13074550 (cit. on p. 11).

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1 (cit. on p. 11).

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis : A survey. https://doi.org/10.48550/arXiv.1801.07883 (cit. on p. 11).

Balakrishnan, V., Shi, Z., Law, C. L., Lim, R., Teh, L. L., & Fan, Y. (2022). A deep learning approach in predicting products' sentiment ratings: A comparative analysis. *The Journal of Supercomputing*, *78*(5), 7206–7226. https://doi.org/10.1007/s11227-021-04169-6 (cit. on p. 11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762 (cit. on p. 11).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805 (cit. on p. 11).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. https://doi.org/10.48550/arXiv.1910.03771 (cit. on p. 12).

Jiang, M., D'Souza, J., Auer, S., & Downie, J. S. (2020). Improving scholarly knowledge representation: Evaluating bert-based models for scientific relation classification. https://doi.org/10.48550/arXiv.2004.06153 (cit. on p. 12).

Huang, S., & Cole, J. M. (2022). Batterybert: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, *62*, 6365–6377. https://doi.org/10.1021/acs.jcim.2c00035 (cit. on p. 12).

Cheng, A., & Gangaraju, S. (2023). Fine-Tuning BERT for Sentiment Analysis, Paraphrase Detection and Semantic Text Similarity. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169889383.pdf (cit. on p. 12).

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Computing Research Repository - CORR*, 417–424. https://doi.org/10.3115/1073083.1073153 (cit. on p. 12).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049 (cit. on p. 12).

Oyebode, O., & Orji, R. (2019). Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019, https://doi.org/{10.1109/IEMCON.2019.8936139} (cit. on p. 13).

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/{10.1609/icwsm.v8i1.14550} (cit. on p. 13).

Al-Shabi, M. (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining, https://api.semanticscholar.org/CorpusID:229705245 (cit. on p. 13).

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (cit. on p. 13).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information (L. Lee, M. Johnson, & K. Toutanova, Eds.). *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051 (cit. on p. 14).

Bloom, B. H. (1970). Space/Time Trade-Offs in Hash Coding with Allowable Errors. *Commun. ACM*, *13*(7), 422–426. https://doi.org/10.1145/362686.362692 (cit. on p. 14).

Boyd, A, & Wamerdam, D. (2022). Floret: Lightweight, robust word vectors. https://explosion.ai/blog/floret-vectors (cit. on p. 14).

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, *47*, 255–308. https://doi.org/10.1162/COLI_a_00402 (cit. on p. 14).

Zúquete, M., Orghian, D., & Pinheiro, F. L. A Moral Foundations Dictionary for the European Portuguese Language: The Case of Portuguese Parliamentary Debates. In: In *Computational Science – ICCS 2023: 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part I*. Prague, Czech Republic: Springer-Verlag, 2023, 421–434. ISBN: 978-3-031-35994-1. https://doi.org/10.1007/978-3-031-35995-8_30 (cit. on p. 27).

Geller, M., Vasconcelos, V. V., & Pinheiro, F. L. Toxicity in evolving twitter topics. In: In *Computational Science – ICCS 2023: 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part IV*. Prague, Czech Republic: Springer-Verlag, 2023, 40–54. ISBN: 978-3-031-36026-8. https://doi.org/10.1007/978-3-031-36027-5_4 (cit. on p. 27).

A

Appendix

Family of Languages	**Languages**
LATIN	French, Italian, Spanish, Portuguese, Romanian
GERMANIC	English, Dutch, German, Danish, Swedish
SLAVIC	Bulgarian, Czech, Polish, Slovak, Slovene
FINNO-UGRIC	Finnish, Hungarian, Estonian
BALTIC	Lithuanian, Latvian
HELLENIC	Greek

Table A.1: Classification of languages by family

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa